Inferring Evolutionary Trees from Ordinal Data*

Paul E. Kearney[1], Ryan B. Hayward[2] and Henk Meijer[3]

[1] Department of Computer Science, University of Toronto,
Toronto ON, CANADA M5S 1A4
kearney@cs.toronto.edu

[2] Department of Mathematics and Computer Science, University of Lethbridge,
4401 University Drive West, Lethbridge AB, CANADA T1K 3M4;
e-mail hayward@alpha.uleth.ca

[3] Department of Computing and Information Science, Queen's University,
Kingston ON, CANADA K7L 3N6;
e-mail henk@qucis.queensu.ca

Abstract. In this paper we present four results on the inference of evolutionary trees from ordinal data. An evolutionary tree T, or phylogeny, is an *ordinal representation* of a distance matrix M if

$$d_T(a, b) \leq d_T(c, d) \text{ if and only if } M[a, b] \leq M[c, d],$$

for all species a, b, c and d under consideration. In particular, we show that

1. Ordinal representations of distance matrices can be found in $O(n^3)$ time where n is the number of species. Ordinal representations are shown to be unique, when they exist.
2. Determining if there is an ordinal representation for an incomplete distance matrix, a situation which arises in evolutionary studies, is NP-complete.
3. Finding an ordinal representation which best fits a distance matrix containing ordinal errors is NP-complete.
4. Under reasonable conditions, a weighted ordinal representation of a distance matrix can be obtained in polynomial time.

1 Inferring Evolutionary Trees from Distance Data

1.1 The Problem of Constructing Evolutionary Trees

A fundamental problem of computational biology is the inferring of evolutionary trees, or phylogenies, from interspecies similarity information. Similarity information usually comes in one of two forms. *Character data* associates species by shared traits such as the possession of a common protein. Shared traits are then interpreted as evidence of shared lineage. *Distance data* is typically derived from experiments in which genetic sequences of two species are aligned and compared. The dissimilarity of the two species is then proportional to the number of sites in which their sequences disagree. This information is catalogued in an $|S| \times |S|$ distance matrix M whose columns and rows are indexed by species in S and the value $M[s, t]$ is the measured dissimilarity of species s and t (the larger $M[s, t]$ is, the more dissimilar are s and t).

In this paper we are primarily interested in inferring phylogenies from distance data. We note that character data is often converted to distance data for analysis. Inference methods interpret

* The authors gratefully acknowledge the support of NSERC.

distance data as evidence of interspecies proximity within the underlying phylogeny. If M exactly matches the distance metric of some phylogeny T then M is called *additive* and T can be recovered easily and efficiently [3].

However, distance matrices obtained through experimentation are never additive due to systematic error. Two such sources of error are *homoplasy* and *superimposed changes*. Homoplasy is the sharing of a trait through evolutionary events other than inheritance from a common ancestor. Hence, character data does not provide categorical evidence of shared lineage. Superimposed changes are mutations which occur at the same site along a genetic sequence. As lineages diverge, the probability that mutations reoccur at the same site will increase. Hence, estimates of evolutionary distance based upon sequence comparisons will in fact underestimate the true evolutionary distance. Several methods have been suggested to account for such errors [12, 15]. Nevertheless, the introduction of error into M has a profound affect on our ability to efficiently and confidently, reconstruct phylogenies from distance data.

1.2 Traditional vs. Ordinal Methods of Phylogenetic Inference

Several techniques for inferring phylogenies from distance data have been proposed. A recent survey of these is given by Swofford et. al [16]. Many of these techniques search for a phylogeny T which minimizes some error function E_T on T and the distance matrix M. For example the error function

$$E_T = \sum_{i<j} w_{ij} |d_T(i,j) - M[i,j]|^\alpha$$

has been studied in [2, 6, 7, 8, 14].

When $\alpha = 1$ the sum of absolute differences is minimized whereas when $\alpha = 2$ a least squares minimization occurs. The weights w_{ij} may be chosen to normalize or possibly to emphasize certain errors. Several NP-Hardness results have been established for finding a phylogeny which minimizes E_T and other error measures [4, 5].

As an alternative to exhaustively searching for a phylogeny which minimizes E_T, heuristic methods have been proposed. Among these are branch-and-bound methods [10, 15], greedy clustering methods as in UPGMA (Unweighted Pair Group Method using Arithmetic averages), the neighbor–joining method [13] and split decomposition [1].

In this paper we examine inference methods which utilize ordinal aspects of distance data. Whereas traditional methods emphasize the representation of *value* (for example, by minimizing E_T), ordinal methods emphasize representation of *relative value*.

More specifically, a phylogeny T is an *ordinal representation* of the distance matrix M, if for all species a, b, c and d,

$$M[a,b] \leq M[c,d] \quad \text{if and only if} \quad d_T(a,b) \leq d_T(c,d)$$

where d_T is the distance metric of T. If T is unweighted then $d_T(x,y)$ is the number of edges on the path $P_T(x,y)$ and if T's edges are positively weighted then $d_T(x,y)$ is the sum of the weights of edges on $P_T(x,y)$. In the latter case T is called a *weighted ordinal representation* of M.

Traditional methods make the assumption that measured dissimilarities correspond, literally, to phylogenetic distances. The existence of superimposed changes suggests that this assumption may be too strong. Ordinal representations assume only that measured distances are indicators of relative species similarity, hence, are more tolerant of errors due to superimposed changes.

1.3 Previous Work on Ordinal Methods

Many established methods for inferring phylogenies employ ordinal methods in some form. For example, the UPGMA method constructs phylogenies by joining the two closest species. These species are then collapsed into a cluster and distances from all other species to this cluster are calculated. The process is then repeated on the next closest pair of species. Hence, the UPGMA method is, in a sense, a greedy ordinal method.

Split decomposition [1] and the Sattath–Tversky method [14] can also be viewed as ordinal methods. A distance matrix M satisfies the *four–point condition* if for all distinct species a, b, c and d, the larger two of the three sums

$$M[a,b] + M[c,d], \quad M[a,c] + M[b,d], \text{ and } M[a,d] + M[b,c]$$

are equal. It is well-known that M is additive exactly when it satisfies the four–point condition [9]. Since experimental data never satisfies the equality constraints of the four–point condition, split decomposition and the Sattath–Tversky method attempt to satisfy relaxations of the four–point condition.

In split decomposition the condition is weakened to the requirement that if $d_T(a,b) + d_T(c,d) \leq d_T(a,c) + d_T(b,d) = d_T(a,d) + d_T(b,c)$ in the true phylogeny T then it is reasonable to assume that $M[a,b] + M[c,d]$ is not larger than $M[a,c] + M[b,d]$ and $M[a,d] + M[b,c]$.

In the Sattath–Tversky method, a measure $s(x,y)$ of how "neighborly" species x and y are is determined based on the following relaxation:

$$s(x,y) = |\{(a,b) \mid M[x,y] + M[a,b] < M[a,x] + M[b,y] \text{ and } M[x,y] + M[a,b] < M[a,y] + M[b,x]\}|.$$

The first explicit study of ordinal methods for inferring phylogenies was undertaken by Kannan and Warnow [11]. In their paper they present an efficient algorithm which reconstructs binary phylogenies from ordinal data and they also establish the NP-completeness of reconstructing phylogenies from partial orders on interspecies similarities.

In this paper we extend the work initiated by Kannan and Warnow by examining both theoretically and biologically motivated problems surrounding ordinal representations. A motivation for studying ordinal representations, and in particular the problems summarized in the overview below, is a lesson learned from the study of additive distance matrices.

A working assumption of phylogenetic inference is that distance matrices which arise from experimentation are nearly additive. Hence, the study of additive matrices, though not directly applicable to the reconstruction of evolutionary trees (because of systematic error), is of extreme importance for establishing a theoretical basis from which efficient approximation algorithms can be developed. In fact, the theory of additive distance matrices has influenced most major methods for inferring phylogenies including UPGMA, neighbor–joining, split decomposition, etc..

As with additivity, not all distance matrices will be ordinal in that they will not have ordinal representations. However, the investigation of those distance matrices which are ordinal and of ordinal inference techniques in general, establishes a basis for the development of applied approximation algorithms for inferring phylogenies.

1.4 Overview of Results

There are four main results which we summarize below.

We present an algorithm which finds an ordinal representation of a given distance matrix in $O(n^3)$ time where n is the number of species (section 3). This extends a result in [11] and requires an in depth analysis of the underlying combinatorics of ordinal representations. As a byproduct of our analysis we establish the uniqueness of ordinal representations which generalizes the result that no two phylogenies have the same distance matrix [9].

We establish the NP-completeness of two biologically relevant problems (section 4). Firstly, often it is the case that distance matrices are incomplete due to error, time and/or cost constraints. We prove that determining if an *incomplete* distance matrix has a weighted ordinal representation is NP-complete.

Secondly, omission of data is not the only type of error which may occur. The distance matrix M may also contain *ordinal errors* where $M[s,x] \leq M[s,y]$ but $d_T(s,y) < d_T(s,x)$ in the true phylogeny T. We demonstrate that the problem of finding good approximate weighted ordinal representations, under a reasonable ordinal error measure, is also NP-complete.

Finally, we show how methods for finding ordinal representations of distance matrices can be extended to derive weighted ordinal representations as well (section 5).

2 Preliminary Definitions

Definition 1. A *phylogeny* T on species set S is a tree (i.e. a connected and acyclic graph) such that species of S are identified with leaves of T and T has no degree two vertices.

It is necessary to require that T have no degree two vertices otherwise edge weights could be simulated by bisecting edges. From a biological perspective, internal nodes of a phylogeny indicate points of speciation, and so, must be of degree at least three. We say T is a *partial* phylogeny if T is permitted to have degree two vertices.

A *subphylogeny* T' of T is a subtree of T which is a phylogeny on some $S' \subseteq S$. In particular, $T[S']$ denotes the subphylogeny *induced* by $S' \subseteq S$. If T is *rooted* at a vertex r then this is denoted by (T,r). An *ultrametric* is a rooted phylogeny where all species are equidistant from the root. The *depth* of an ultrametric is the distance from any species to the root.

The *contraction* of T at an edge pq is the phylogeny which results from identifying nodes p and q. An *expansion* of T at a node p is the inverse operation of a contraction. We say M is the distance matrix of the phylogeny T if $M[a,b] = d_T(a,b)$, for all $a,b \in S$.

3 An $O(n^3)$ Algorithm for Inferring Phylogenies from Ordinal Data

In this section we detail an algorithm which finds an ordinal representation of a distance matrix M over species set S, if one exists. The algorithm breaks naturally into two stages. In the first stage a *midpath phylogeny* (defined below) for M is determined which provides us with an approximate solution. The second stage extends the midpath phylogeny to an ordinal representation of M. An advantage of this division is that the midpath phylogeny can also be used to find a weighted ordinal representation of M. This will be discussed in section 5.

3.1 Midpath Trichotomy

Let T be a phylogeny on S and M a distance matrix over S. Define the *midpoint* of $P_T(x,y)$, $mid(x,y)$, to be that vertex or edge on $P_T(x,y)$ equidistant from x and y. With respect to M, define the *midpath* of $P_T(x,y)$, $m(x,y)$, to be that vertex or edge on $P_T(x,y)$, if it exists, such that the following *trichotomy property* holds:

- All species s in the connected component of $T - \{m(x,y)\}$ containing x satisfy $M[s,x] < M[s,y]$.
- All species s in the connected component of $T - \{m(x,y)\}$ containing y satisfy $M[s,y] < M[s,x]$.
- All species s in other connected components of $T - \{m(x,y)\}$ satisfy $M[s,x] = M[s,y]$.

For example, in Figure 1, the midpoint and midpath for species x and y are indicated.

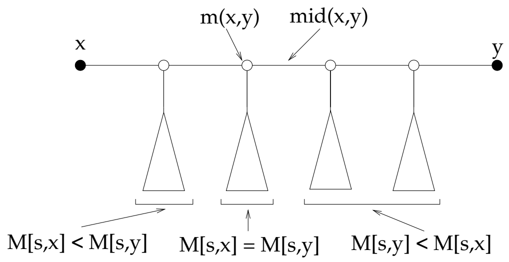

Fig. 1. Trichotomy of S induced by $m(x,y)$.

Clearly, if T is an ordinal representation for M then $m(x,y) = mid(x,y)$ for all pairs of species x and y. Hence, a necessary condition for ordinal representation is that all pairs of species have midpaths. This motivates the following definition:

Definition 2. T is a *midpath phylogeny* for M if T is a phylogeny on S such that midpaths exist for all species pairs. If, in addition, no contraction of T is a midpath phylogeny for M then T is called *minimal.*

Intuitively, the existence of a midpath phylogeny for M indicates that M is "tree-like" in that all midpaths can be aligned on a common phylogeny — the ordinality of M is consistent with a phylogenetic structure.

3.2 Essential Properties of Midpath Phylogenies

Property 3 *If it exists, the minimal midpath phylogeny of M is unique.*

Proof. Omitted. □

Algorithmically, the most useful property of minimal midpath phylogenies is the following:

Theorem 4. *Any ordinal representation of M is an expansion of the minimal midpath phylogeny of M.*

Proof. Omitted. □

We have developed an optimal $O(n^2)$ algorithm which constructs the minimal midpath phylogeny of M, if it exists. To achieve the optimal running time midpath trichotomy constraints must be successively satisfied in a carefully chosen order dependent upon the ordinal structure of M. Here we omit details on how this is accomplished.

3.3 Expanding the Minimal Midpath Phylogeny to an Ordinal Representation

According to Theorem 4, to obtain an ordinal representation T_R of M we must expand the minimal midpath phylogeny T of M. Intuitively this task can be equated to expanding T to a phylogeny T_R so that the midpath $m(x, y)$ and midpoint $mid(x, y)$ are the same, for all species x and y in S.

We provide an overview of the expansion algorithm. Due to the complexity of the key results, their proofs are included in the appendix.

Root T at any internal vertex r and consider the locality of a vertex v as depicted in Figure 2. The vertices v_1, v_2, ..., v_m are the *children* of v. $T(v)$ denotes the subphylogeny of T rooted at v and $S(v)$ denotes the set of species in $T(v)$.

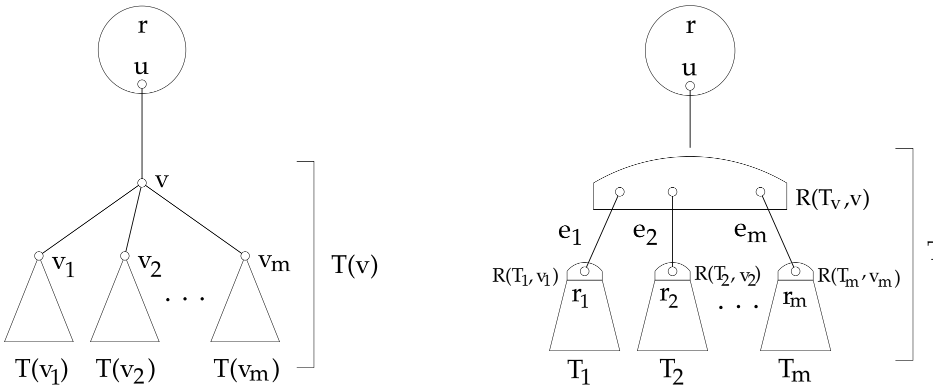

Fig. 2. The locality of v in the rooted minimal midpath phylogeny T and the expansion $T_v \in Exp(v)$ of $T(v)$ and its components.

T_R is assembled from subsolutions. Hence, consider the conditions a subsolution $T_v = T_R[S(v)]$ must satisfy:

Condition 1: T_v is an expansion of $T(v)$ since T_R is an expansion of T by Theorem 4.

Condition 2: T_v has at most one degree two vertex and it must be located in $R(T_v, v)$ where $R(T_v, v)$ denotes the expansion of v in T_v. We must permit a degree two vertex in $R(T_v, v)$ since in T_R there is an edge pq connecting T_v to $T_R - T_v$ where $p \in R(T_v, v)$. Conversely, since T_R cannot have degree two vertices, there can be no more than one degree two vertex in T_v.

Condition 3: T_v is an ordinal representation of the submatrix of M induced by species in $S(v)$ since T_R is an ordinal representation of M.

Condition 4: For all $x, y \in S(v)$, if $m(x, y) = vv_i$, for some i, then $d_{T_v}(x, e_i) = d_{T_v}(y, e_i)$ where e_i is the edge of T_v connecting $R(T_v, v)$ and $R(T_v, v_i)$. If $m(x, y) = v$ then $d_{T_v}(x, e_i) = d_{T_v}(y, e_j)$. These constraints are necessary since $m(x, y)$ specifies where the midpoint $mid(x, y)$ should be located in T_v.

Definition 5. $Exp(v)$ is the set of all phylogenies on $S(v)$ satisfying the above four conditions.

Our algorithm proceeds by determining $Exp(v)$ for each vertex v of T, and so, $T_R \in Exp(r)$. $Exp(v)$ is determined dynamically from $Exp(v_1)$, $Exp(v_2)$, ..., $Exp(v_m)$. If $T_v \in Exp(v)$ then T_v consists of $R(T_v, v)$ and the rooted subphylogenies (T_1, r_1), (T_2, r_2), ..., (T_m, r_m) where $T_i \in Exp(v_i)$ and $r_i \in R(T_i, v_i)$, $1 \leq i \leq m$. T_v and its parts are depicted in Figure 2.

In general, $Exp(v_i)$ may not be a singleton set nor is $R(T_i, v_i)$ always a single vertex. Consequently, determining $Exp(v)$ by performing a brute-force search on all combinations of rooted subphylogenies (T_1, r_1), (T_2, r_2), ..., (T_m, r_m) would be inefficient.

In the remainder of this section we present several structural and algorithmic results. In particular, we introduce and develop the concept of a *switch* which is a pair of ordinally equivalent[4] phylogenies. Switches play a key role in the expansion algorithm and in the underlying combinatorics of ordinal representations. The key result being that $Exp(v)$ is either uniquely determined or is a switch (the One–Two Theorem).

Subsolution Selection: Suppose $(T_1, r_1), \ldots, (T_m, r_m)$ are rooted subphylogenies of $T_v \in Exp(v)$ where $T_i \in Exp(v_i)$ and $r_i \in R(T_i, v_i)$, $1 \leq i \leq m$. Then the values $d_{T_v}(r_i, r_j)$, for $1 \leq i < j \leq m$, can be determined using *midpath tests*. Although they are not detailed here, midpath tests use the location of midpaths in T to deduce the values $d_{T_v}(r_i, r_j)$, for $1 \leq i < j \leq m$.

A *unique* phylogeny T' with leaves r_1, r_2, ..., r_m which realizes the distances $d_{T'}(r_i, r_j) = d_{T_v}(r_i, r_j)$ can be found in $\Theta(m^2)$ time, if one exists, using additive tree reconstruction methods [3].

We denote the tree obtained by identifying the leaves r_1, r_2, ..., r_m of T' with the roots r_1, r_2, ..., r_m of T_1, T_2, ..., T_m by $rec((T_1, r_1), (T_2, r_2), \ldots, (T_m, r_m))$. (T_i, r_i) is called a *component* of $rec((T_1, r_1), (T_2, r_2), \ldots, (T_m, r_m))$. From these definitions we have,

$$T_v = rec((T_1, r_1), (T_2, r_2), \ldots, (T_m, r_m)).$$

Definition 6. A (n, k)-*switch* S is a pair of trees $S[ON]$ and $S[OFF]$ each composed of $n+1$ ultrametrics U_1, U_2, ..., U_n and U' all of depth k along with a rooted phylogeny U. U and U' are connected at their roots by a path w_0, w_1, ..., w_{n+1} of length $n+1$ where w_0 is the root of U and w_{n+1} is the root of U'. $U \cup \{w_0, w_1, \ldots, w_{n+1}\} \cup U'$ is called the *spine* of the switch. Furthermore,

- In $S[ON]$ the roots of U_1, U_2, ..., U_n are identified with the vertices w_1, w_2, ..., w_n, respectively.
- In $S[OFF]$ the roots of U_1, U_2, ..., U_n are adjacent to the vertices w_0, w_1, ..., w_{n-1}, respectively.
- Finally, $S[ON]$ and $S[OFF]$ are ordinally equivalent.

So that we may refer to particular components of a switch S we write $S = \{U, U_1, U_2, \ldots, U_n, U', w_{n+1}, w_n\}$.

Switches are best described by pictures. In Figure 3 all species are located at the leaves of the subphylogenies U_1, U_2, U' and U. U_1, U_2, and U' are ultrametrics of depth 4 whereas U is an ultrametric of depth 1. Observe that the only interspecies distances which differ between the pair of phylogenies are distances between pairs of leaves taken from U_1, U_2 or U'. However, these distances increase uniformly from $S[ON]$ to $S[OFF]$ and are strictly larger than the distances which do not change. Hence, $S[ON]$ and $S[OFF]$ are ordinally equivalent.

Joining Lemma. *If $T_v, T'_v \in Exp(v)$ where $T_v = rec((T_1, r_1), \ldots, (T_m, r_m))$, $T'_v = rec((T_1, r'_1), \ldots, (T_m, r'_m))$ and $r_i \neq r'_i$, for some i then*

[4] Two phylogenies T_1 and T_2 on the same species set S are ordinally equivalent if, for all species a, b, c and d in S, $d_{T_1}(a, b) \leq d_{T_1}(c, d)$ if and only if $d_{T_2}(a, b) \leq d_{T_2}(c, d)$.

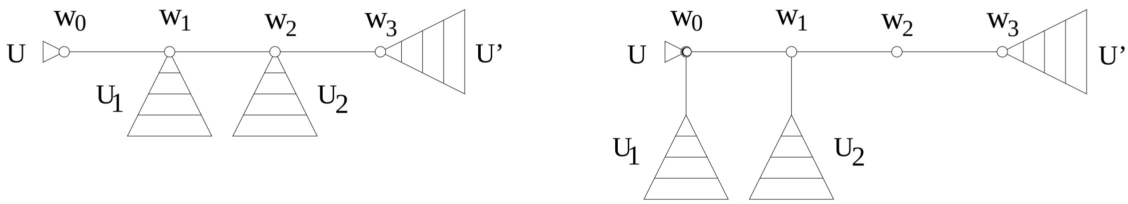

Fig. 3. A $(2,4)$-switch $S = (U, U_1, U_2, U', w_3, w_2)$ where $S[ON]$ is on the left and $S[OFF]$ is on the right.

- $r_j = r'_j$, $1 \leq j \leq m$, $j \neq i$,
- $d(r_i, r'_i) = 1$,
- $T' = rec((T_1, r_1), \ldots, (T_{i-1}, r_{i-1}), (T_{i+1}, r_{i+1}), \ldots, (T_m, r_m))$ *is an ultrametric of some depth* k *and*
- $\{T_v, T'_v\}$ *is a* $(2, k)$*-switch, for some* p.

Proof. Please see the appendix for full details. □

Our main result is that $Exp(v)$ is either unique or a switch. Hence, as the algorithm progresses the number of subsolutions remains small.

One–Two Theorem. *If* $Exp(v) \neq 0$ *then Either* $|Exp(v)| = 1$ *or* $Exp(v)$ *is a switch.*

Proof. Please see the appendix for full details. □

3.4 Uniqueness

As a corollary to the One–Two Theorem, if M has an ordinal representation T_R then T_R is unique. This follows from the One–Two Theorem since either $|Exp(r)| = 1$, hence uniqueness, or $Exp(r)$ is a switch T_r. However, $T_r[OFF]$ contains a degree two vertex, hence, $T_r[OFF]$ cannot be an ordinal representation of M.

This generalizes the well-known result [9] that distance matrices have unique additive representations. In fact, distance matrices have unique ordinal representations.

4 Two Complexity Results

In this section we define and discuss the biological relevance of two complexity results.

4.1 Incomplete Data

Often it is the case that evolutionary surveys of interspecies similarities are not complete due to error, time and/or cost constraints. In this case we must derive an ordinal representation from an *incomplete* matrix — a matrix where some entries are missing.

Ordinal Representation from Incomplete Matrix (ORIM):
Instance: A distance matrix M over species set S where M may be incomplete.
Problem: Is there a weighted phylogeny T on species set S such that, for all existing values $M[a,b]$ and $M[c,d]$, $M[a,b] \leq M[c,d]$ if and only if $d_T(a,b) \leq d_T(c,d)$?

Theorem 7. **ORIM** *is NP-complete.*

Proof. Omitted. □

This result contrasts with the NP-completeness result of Kannan and Warnow [11]. They show that if not all comparisons among interspecies similarities are known then determining if an ordinal representation exists is NP-complete. However, if comparative information is derived from an incomplete distance matrix then all comparisons among existing values $M[a,b]$ are necessarily known and cannot be disregarded. Hence, our formulation of absent information and that found in [11] differ significantly.

4.2 Ordinal Errors

Distance matrices arising from experimental data may contain ordinal errors. The pair of species x and y form an *ordinal error* if $M[s,x] \leq M[s,y]$ but in the true phylogeny T, $d_T(s,y) < d_T(s,x)$, for some species s.

In an effort to recover the true phylogeny, we seek a weighted phylogeny T which minimizes the number of ordinal errors with M.

Best Ordinal Representation (BOR):
Instance: A distance matrix M over species set S and a positive integer k.
Problem: Is there a weighted phylogeny T on species set S such that $E(T,M) \leq k$ where $E(T,M)$ is the number of species pairs x and y such that $M[s,x] \leq M[s,y]$ but $d_T(s,y) < d_T(s,x)$, for some species s?

Theorem 8. **BOR** *is NP-complete.*

Proof. Omitted. □

5 Weighted Ordinal Representations

Methods used to determine the ordinal representation of a distance matrix M can also be used to determine a weighted ordinal representation of M. In particular, Theorem 4 applies to weighted ordinal representations: A weighted ordinal representation of M is an expansion of the minimal midpath phylogeny of M.

However, the discrete methods used in the expansion algorithm (section 3.3) cannot be extended to the weighted case. Nevertheless, for any expansion T of the minimal midpath phylogeny T_M, a linear inequality program can be defined to determine if T's edges can be weighted so that T is a weighted ordinal representation of M. Such programs can be solved in polynomial time.

The difficulty is that the minimal midpath phylogeny may not be "well-developed" in the sense that there may be an exponential number of expansions of T_M. In this case finding an expansion of T_M which forms a weighted ordinal representation of M takes an exponential amount of time.

Under the assumption that the minimal midpath phylogeny is within a constant number of expansions of being binary[5] then the brute-force search algorithm is polynomial. How reasonable this assumption is depends upon how well articulated the true evolutionary tree is with respect to ordinality and, if it is, how well the experimental methods used to derive M detect this articulation.

References

1. H.-J. Bandelt and A. W. M. Dress. Split decomposition: A new and useful approach to phylogenetic analysis of distance data. *Mol. Phyl. Evol.* **1**, 1992.
2. L. L. Cavalli–Sforza and A. W. F. Edwards. Phylogenetic analysis: Models and estimation procedures. *Evolution* **32**, 1967.
3. J. C. Culberson and P. Rudnicki. A fast algorithm for constructing trees from distance matrices. *Information Processing Letters,* 30:215 – 220, 1989.
4. William H. E. Day. Inferring phylogenies from dissimilarity matrices. *Bull. of Math. Biol.* **49**:4, 1987.
5. M. Farach, S. Kannan and T. Warnow. A robust model for finding evolutionary trees. *Algorithmica* **13**:1, 1995.
6. J. S. Farris. Estimating phylogenetic trees from distance matrices. *Am. Nat.* **106**, 1972.
7. J. Felsenstein. *PHYLIP (Phylogeny Inference Package)*, version 3.5c. Department of Genetics, University of Washington, Seattle, 1993.
8. W. M. Fitch and E. Margoliash. Construction of phylogenetic trees. *Science* **155**, 1967.
9. S. L. Hakimi and S. S. Yau. Distance matrix of a graph and its realizability. *Quarterly of Applied Mathematics,* 22:305 – 317, 1964.
10. M. D. Hendy and D. Penny. Branch and bound algorithms to determine minimal evolutionary trees. *Math. Biosci.* **59**, 1982.
11. S. Kannan and T. Warnow. Tree reconstruction from partial orders. *SIAM J. Comp.*, **24**:3, 1995.
12. S. Kumar, K. Tamura and M. Nei. *MEGA: Molecular Evolutionary Genetics Analysis. Version 1.0.* Pennsylvania State University, University Park, Pennsylvania, 1993.
13. N. Saitou and M. Nei. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 1987.
14. S. Sattath and A. Tversky. Additive similarity trees. *Psychometrika* **42**, 1977.
15. David L. Swofford. *PAUP*: Phylogenetic Analysis Using Parsimony (and Other Methods)*, version 4.0. Sinauer Associates, Sunderland, Massachusetts, 1996.
16. David L. Swofford, Gary J. Olsen, Peter J. Waddell and David M. Hillis. Phylogenetic Inference. In *Molecular Systematics*, 2nd Edition. David M. Hillis, Craig Moritz and Barbara K. Mable (eds.), Sinauer Associates, Sunderland, Massachusetts, 1996.

[5] T is binary if all internal vertices have degree three.

Appendix: Technical Developments

The primary results which follow are the Joining Lemma and the One–Two Theorem presented in section 3.3.

Shallowest Depth Lemma. *A rooted phylogeny where only the root is permitted to have degree two and the shallowest depth is k has species pairs which are distances $0, 2, \ldots, 2k$ apart.*

Proof. Omitted. □

Joining Lemma. *If $T_v, T_v' \in Exp(v)$ where $T_v = rec((T_1, r_1), \ldots, (T_m, r_m))$, $T_v' = rec((T_1, r_1'), \ldots, (T_m, r_m'))$ and $r_i \neq r_i'$, for some i then*

- $r_j = r_j'$, $1 \leq j \leq m$, $j \neq i$,
- $d(r_i, r_i') = 1$,
- $T' = rec((T_1, r_1), \ldots, (T_{i-1}, r_{i-1}), (T_{i+1}, r_{i+1}), \ldots, (T_m, r_m))$ *is an ultrametric of some depth k and*
- $\{T_v, T_v'\}$ *is a $(2, k)$-switch, for some p.*

Proof. Let $j \neq i$ and consider how T_i and T_j are connected in T_v and in T_v'. Let $d_1 = d_{T_v}(r_i, r_j)$, $d_2 = d_{T_v'}(r_i', r_j')$, $s = d_{T_i}(r_i, r_i')$ and $t = d_{T_j}(r_j, r_j')$. Let $S_a \subseteq S(v_i)$ be those species s such that $r_i \in P(s, r_i')$. Let a be the shallowest species in the rooted subphylogeny $(T_i[S_a], r_i)$. Let $S_b \subseteq S(v_i)$ be those species s such that $r_i' \in P(s, r_i)$. Let b be the shallowest species in the rooted subphylogeny $(T_i[S_b], r_i')$. Let $S_c \subseteq S(v_j)$ be those species s such that $r_j \in P(s, r_j')$. Let c be the shallowest species in the rooted subphylogeny $(T_j[S_c], r_j)$. Let $S_d \subseteq S(v_j)$ be those species s such that $r_i' \in P(s, r_i)$. Let d be the shallowest species in the rooted subphylogeny $(T_j[S_d], r_j')$. This is depicted in Figure 4.

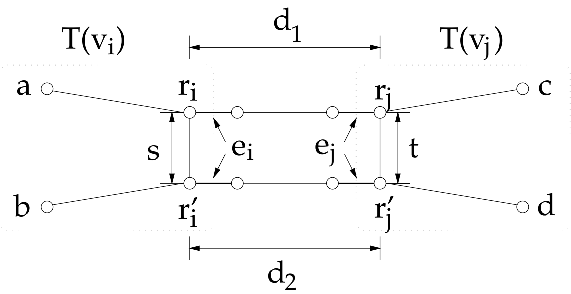

Fig. 4. How $T(v_i)$ and $T(v_j)$ join in T_v and T_v'.

Dependencies among the values d_1, d_2, s and t can be obtained by considering the locations of midpaths $m(a, c)$, $m(a, d)$, $m(b, c)$ and $m(b, d)$ in the minimal midpath phylogeny T. For example, if $m(a, c) = vv_i$ then since T_v and $T_v' \in Exp(v)$ satisfy Condition 4, $d_{T_v}(a, e_i) = d_{T_v}(c, e_i)$ and $d_{T_v'}(a, e_i) = d_{T_v'}(c, e_i)$. It follows that $d_{T_i}(a, r_i) = d_1 - 1 + d_{T_j}(c, r_j)$ and $d_{T_i}(a, r_i) + s = d_2 - 1 + t + d_{T_j}(c, r_j)$ which implies the dependency $d_1 + s = d_2 + t$.

This technique in conjunction with the fact that any midpath $m(x,y)$ of T, where $x \in T_i$ and $y \in T_j$, must be located in $T(v_i) \cup \{vv_i, v, vv_j\} \cup T(v_j)$ allows us to generate a *midpath survey* (Table 1) of dependencies among d_1, d_2, s and t.

	$m(a,c)$	$m(a,d)$	$m(b,c)$	$m(b,d)$
$T(v_i)$	$d_1 = d_2 + s + t$	$d_1 + t = d_2 + s$	$d_1 + s = d_2 + t$	$d_2 = d_1 + s + t$
vv_i	$d_1 + s = d_2 + t$	$d_2 = d_1 + s + t$	$d_1 = d_2 + s + t$	$d_1 + t = d_2 + s$
v	$d_1 = d_2 = 2$ $s = t$	not consistent	not consistent	$d_1 = d_2 = 2$ $s = t$
vv_j	$d_1 + t = d_2 + s$	$d_1 = d_2 + s + t$	$d_2 = d_1 + s + t$	$d_1 + s = d_2 + t$
$T(v_j)$	$d_1 = d_2 + s + t$	$d_1 + t = d_2 + s$	$d_1 + s = d_2 + t$	$d_2 = d_1 + s + t$

Table 1. Midpath location and the resulting relation on d_1, d_2, s and t.

Observe that either $d_{T_i}(a, r_i) \geq d_{T_i}(b, r_i') + s$ or $d_{T_i}(b, r_i') \geq d_{T_i}(a, r_i) + s$ since otherwise $d_{T_v}(a,c) < d_{T_v}(b,c)$ and $d_{T_v'}(b,c) < d_{T_v'}(a,c)$ contradicting both T_v and T_v' satisfying Condition 3. Without loss of generality, assume $d_{T_i}(a, r_i) \geq d_{T_i}(b, r_i') + s$.

We proceed with a case-by-case analysis based upon the location of $m(b,c)$ in T:

Case $m(b,c) \in T(v_i)$**:**

By the midpath survey, $d_1 + s = d_2 + t$. Since $m(b,c) \in T(v_i)$ and $d_{T_i}(a, r_i) \geq d_{T_i}(b, r_i') + s$, it must be that $m(a,c) \in T(v_i)$. By the midpath survey, $d_1 = d_2 + s + t$ which implies $s = 0$ contradicting $r_i \neq r_i'$.

Case $m(b,c) = vv_i$**:**

By the midpath survey, $d_1 = d_2 + s + t$. Consider the possible locations for $m(b,d)$. If $m(b,d) \in T(v_i)$ then $d_2 = d_1 + s + t$, hence, $s = 0$ — a contradiction. Similarly, if $m(b,d) \in T(v_j) \cup \{v, vv_j\}$ then $s = 0$. It follows that $m(b,d) = vv_i$ which implies $t = 0$ (i.e. $r_j = r_j'$) and $d_1 = d_2 + s$ by the midpath survey, and so, for all $b \in T_i[S_b]$ and all $c \in T_j[S_c]$ (note that $T_j[S_c] = T_j[S_d]$ since $t = 0$), $m(b,c) = vv_i$.

By Condition 4, $T_i[S_b]$ and $T_j[S_c]$ are ultrametrics of depth k and $k - d_2 + 1$ respectively, for some k.

Case $m(b,c) = v$**:**

By the midpath survey, this is an impossibility since it necessitates an inconsistent dependency among d_1, d_2, s and t.

Case $m(b,c) = vv_j$**:**

This case is analogous to the case $m(b,c) = vv_i$.

Case $m(b,c) \in T(v_j)$**:**

By the midpath survey, $d_1 + s = d_2 + t$. Consider the possible locations for $m(a,d)$. If $m(a,d) \in \{vv_i, vv_j\}$ then the midpath survey informs us that $s = 0$ — a contradiction. Since $m(a,d) = v$ is inconsistent, it follows that $m(a,d) \in T(v_i) \cup T(v_j)$.

If $m(a,d) \in T(v_j)$ then $m(b,d) \in T(v_j)$ since $d_{T_i}(a,r_i) \geq d_{T_i}(b,r_i') + s$, and by the midpath survey, $t = 0$. It follows that the only noncontradictory location for $m(a,c)$ is vv_i. Hence, $d_1 + s = d_2$ and for all $a \in T_i[S_a]$, and all $c \in T_j[S_c]$, $m(a,c) = vv_i$. This implies that $T_i[S_a]$ is an ultrametric of depth k and $T_j[S_c]$ is an ultrametric of depth $k - d_1 + 1$, for some k.

If $m(a,d) \in T(v_i)$ then $s = t > 0$ and $d_1 = d_2$ by the midpath survey. Consider how $d(b,d)$ varies between T_v and T_v'. In the former, $d_{T_v}(b,d) = d_{T_i}(b,r_i') + s + d_1 + t + d_{T_j}(d,r_j')$ and in the latter $d_{T_v'}(b,d) = d_{T_i}(b,r_i') + d_2 + d_{T_j}(d,r_j')$. Notice that $d_{T_v}(b,d) > d_{T_v'}(b,d)$ since $s = t > 0$ and $d_1 = d_2$. Since $m(a,d) \in T(v_i)$, it follows that $d_{T_i}(a,r_i) \geq d_{T_j}(d,r_j') + d_1 + t$. Combining this with $d_{T_i}(a,r_i) \geq d_{T_i}(b,r_i') + s$ yields

$$2d_{T_i}(a,r_i) \geq d_{T_i}(b,r_i') + s + d_1 + t + d_{T_j}(d,r_j'),$$

hence, $2d_{T_i}(a,r_i) \geq d_{T_v}(b,d) > d_{T_v'}(b,d)$. By definition a is the shallowest species in $T_i[S_a]$, and so by the Shallowest Depth Lemma, $T_i[S_a]$ has species pairs distances $0, 2, \ldots, 2d_{T_i}(a,r_i)$ apart. It follows that there exists species pairs uv and wx in $T_i[S_a]$ such that $uv < bd \leq wx$ in M and $d_{T_i}(u,v) = d_{T_i}(w,x) - 2$. This implies that $d(b,d)$ may have exactly one value, and so, not both T_v and T_v' are elements of $Exp(v)$.

Our case-by-case analysis has shown that, for all $j \neq i$, $r_j = r_j'$ and either $d_2 = d_1 + s$, $T_i[S_a]$ is an ultrametric of depth k and T_j is an ultrametric of depth $k - d_1 + 1$, for some k, or $d_1 = d_2 + s$, $T_i[S_b]$ is an ultrametric of depth k and T_j is an ultrametric of depth $k - d_2 + 1$, for some k.

We now show that $s = 1$. Assume $s > 1$, hence, there exists some $p \in P_{T_i}(r_i, r_i')$ such that p is adjacent to r_i. Since $P_{T_i}(r_i, r_i') \not\subseteq R(T_v, v)$, p is not a degree two vertex by Condition 2. It follows that there exists a species x which inserts onto $P_{T_i}(r_i, r_i')$ at p. Using the same technique to derive the midpath survey of Table 1 we can derive relations on d_1, d_2 and s by considering possible locations for $m(c,x)$:

If $m(c,x) \in T(v_i) \cup \{vv_j\} \cup T(v_j)$ then $d_1 + 2 = d_2 + s$. If $d_1 = d_2 + s$ then a contradiction occurs. Otherwise, $d_2 = d_1 + s$ which implies $s = 1$ contradicting the assumption $s > 1$.

If $m(c,x) = vv_i$ then $d_1 + s = d_2 + 2$. If $d_1 = d_2 + s$ then $s = 1$ contradicting $s > 1$. Otherwise, $d_2 = d_1 + s$ which also results in a contradiction.

Finally, if $m(c,x) = v$ then $d_2 = d_1$ which contradicts $d_2 = d_1 + s$ and $d_1 = d_2 + s$ since $s > 1$. We conclude that $s = 1$.

To complete the proof, define $T' = rec((T_1,r_1), \ldots, (T_{i-1}, r_{i-1}), (T_{i+1}, r_{i+1}), \ldots, (T_m, r_m))$. Hence, $T_i = T_v - T' = T_v' - T'$. Let $P_{T_v}(r_i, q)$ be the path connecting T_i and T' in T_v and let $P_{T_v'}(r_i', q')$ be the path connecting T_i and T' in T_v'.

The above case-by-case analysis which analyzed how T_i and T_j join in T_v and T_v' can be repeated to analyze how T_i and T' are joined in T_v and T_v' where T_j is replaced by T', r_j is replaced by q, r_j' is replaced by q', $d_1 = d_{T_v}(r_i, q)$ and $d_2 = d_{T_v'}(r_i', q')$. As above, we may conclude that $q = q'$ and either $d_1 = d_2 + 1$, $T_i[S_b]$ is an ultrametric of depth k and T' is an ultrametric rooted at q of depth $k - d_2 + 1$, for some k, or, $d_2 = d_1 + 1$, $T_i[S_a]$ is an ultrametric of some depth k and T' is an ultrametric of depth $k - d_1 + 1$, for some k.

Since T_v and T_v' satisfy Condition 2, there can be at most one degree two vertex on $P_{T_v}(r_i, q)$ and on $P_{T_v'}(r_i', q')$. It follows that either $d_1 = 2$ and $d_2 = 1$, or, $d_1 = 1$ and $d_2 = 2$. We conclude that $\{T_v, T_v'\}$ is either a $(2,k)$-switch $(T_i[S_a], T_i[S_b], T', q, w)$, where w is adjacent to r_i and q in T_v, or, $\{T_v, T_v'\}$ is a $(2,k)$-switch $(T_i[S_b], T_i[S_a], T', q, w)$, where w is adjacent to r_i' and q in T_v' This concludes the proof. $\square$

Switch Root Lemma. *Let $Exp(v_i)$ be a (n,k)-switch $T_i = (U, U_1, U_2, \ldots, U_n, U', r_{ON}, r_{OFF})$.*

For any $T_v \in Exp(v)$ either $T_v = rec((T_1, r_1), \ldots, (T_i[ON], r_{ON}), \ldots, (T_m, r_m))$ or $T_v = rec((T_1, r_1), \ldots, (T_i[OFF], r_{OFF}), \ldots, (T_m, r_m))$.

Proof. Omitted. $\square$

Switch Polarization Lemma. *Let $Exp(v_i)$ be a (p,k)-switch $T_i = \{U, U_1, U_2, \ldots, U_p, U', r_{ON}, r_{OFF}\}$ and $Exp(v_j)$ be a (q,l)-switch $T_j = \{W, W_1, W_2, \ldots, W_q, W', s_{ON}, s_{OFF}\}$. If $T_v, T_v' \in Exp(v)$ where $T_v = rec((T_1, r_1), \ldots, (T_i[x], r_x), \ldots, (T_j[y], r_y), \ldots, (T_m, r_m))$ and $T_v' = rec((T_1, r_1'), \ldots, (T_i[x'], r_{x'}), \ldots, (T_j[y'], r_{y'}), \ldots, (T_m, r_m'))$ then $x = x'$ or $y = y'$.*

Proof. Omitted. $\square$

Switch Joining Lemma. *Let $Exp(v_i)$ be a (q,k)-switch $T_i = (U, U_1, \ldots, U_q, U', r_{ON}, r_{OFF})$. Let $T_v, T_v' \in Exp(v)$ where $T_v = rec((T_1, r_1), \ldots, (T_i[ON], r_{ON}), \ldots, (T_m, r_m))$, $T_v' = rec((T_1, r_1'), \ldots, (T_i[OFF], r_{OFF}), \ldots, (T_m, r_m'))$.*

Letting $T' = rec((T_1, r_1), \ldots, (T_{i-1}, r_{i-1}), (T_{i+1}, r_{i+1}), \ldots, (T_m, r_m))$,

- *$r_j = r_j'$, $1 \leq j \leq m$, $j \neq i$,*
- *T' is an ultrametric of depth k rooted at some vertex r',*
- *$T_i[ON]$, T' are connected by an edge $r_{ON}r'$ in T_v and*
- *$T_i[ON]$, T' are connected by a path $P_{T_v'}(r_{OFF}, r')$ of length two in T_v'.*

Proof. Omitted. $\square$

One–Two Theorem. *If $Exp(v) \neq 0$ then Either $|Exp(v)| = 1$ or $Exp(v)$ is a switch.*

Proof. By induction, $|Exp(v_i)| = 1$ or $Exp(v_i)$ is a switch, $1 \leq i \leq m$.

Case: There are no switches among $Exp(v_1)$, $Exp(v_2)$, …, $Exp(v_m)$.

Let $Exp(v_i) = \{T_i\}$, $1 \leq i \leq m$. If $|Exp(v)| = 1$ then we are done, otherwise, suppose $T_v, T_v' \in Exp(v)$ are distinct where $T_v = rec((T_1, r_1), \ldots, (T_m, r_m))$, $T_v' = rec((T_1, r_1'), \ldots, (T_m, r_m'))$ and $r_i \neq r_i'$, for some i.

By the Joining Lemma, $\{T_v, T_v'\}$ forms a $(1,k)$-switch $(T_i(r_i), T_i(r_i'), T', p, p')$ where $T_i(r_i)$ is the rooted subphylogeny of T_i at r_i', $T_i(r_i')$ is the rooted subphylogeny of T_i at r_i, $T' = rec((T_1, r_1), \ldots, (T_{i-1}, r_{i-1}), (T_{i+1}, r_{i+1}), \ldots, (T_m, r_m))$ is an ultrametric of depth k rooted at a vertex p and p' is adjacent to r_i and p in T_v. This is depicted in Figure 5.

Suppose $\hat{T_v} \in Exp(v) - \{T_v, T_v'\}$ where $\hat{T_v} = rec((T_1, \hat{r_1}), \ldots, (T_m, \hat{r_m}))$. If $r_j \neq \hat{r_j}$, $j \neq i$, then by the Joining Lemma, $r_i = \hat{r_i} = r_i'$, a contradiction. It follows that $\hat{r_i}$ is distinct from r_i and r_i'. The Joining Lemma implies that r_i, r_i' and $\hat{r_i}$ are pairwise adjacent in T_i contradicting the acyclicity of T_i. We conclude that $Exp(v)$ is a switch.

Case: There is one switch among $Exp(v_1)$, $Exp(v_2)$, …, $Exp(v_m)$.

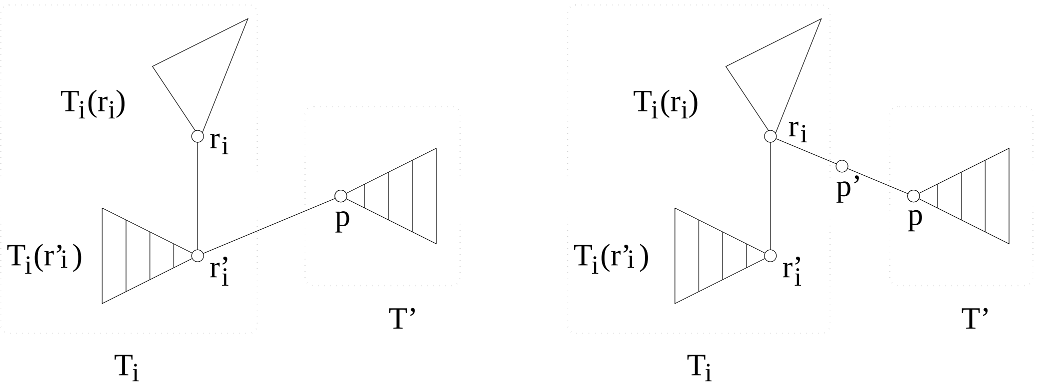

Fig. 5. T'_v and T_v: the creation of a switch.

Let $Exp(v_i)$ be a (q,k)-switch $T_i = (U, U_1, U_2, \ldots, U_q, U', r_{ON}, r_{OFF})$ and let $Exp(v_j) = \{T_j\}$, $1 \le j \le m$, $j \ne i$. If $|Exp(v)| = 1$ then we are done, otherwise, let $T_v, T'_v \in Exp(v)$ where $T_v = rec((T_1, r_1),\ \ldots,\ (T_i[x], r_y),\ \ldots,\ (T_m, r_m))$, $T'_v = rec((T_1, r'_1),\ \ldots,\ (T_i[x'], r_{y'}),\ \ldots,\ (T_m, r'_m))$ and $x, y, x', y' \in \{ON, OFF\}$.

By the Switch Root Lemma, $x = y$ and $x' = y'$. If $x = y$ then since T_v and T'_v are distinct, $r_j \ne r'_j$ for some $j \ne i$. By the Joining Lemma this implies that $T_i[x]$ is an ultrametric which is impossible since T_i is a switch. We conclude that $x \ne y$. Without loss of generality, assume $x = ON$ and $y = OFF$.

By the Switch Joining Lemma, $\{T_v, T'_v\}$ form a $(q+1,k)$-switch $(U, U_1, U_2, \ldots, U_q, U', T', p, p')$ where $T' = rec((T_1, r_1),\ \ldots,\ (T_{i-1}, r_{i-1}),\ (T_{i+1}, r_{i+1}),\ \ldots,\ (T_m, r_m))$ is an ultrametric of depth k with root p and p' is adjacent to r_{OFF} and p in T_v. This is depicted in Figure 6.

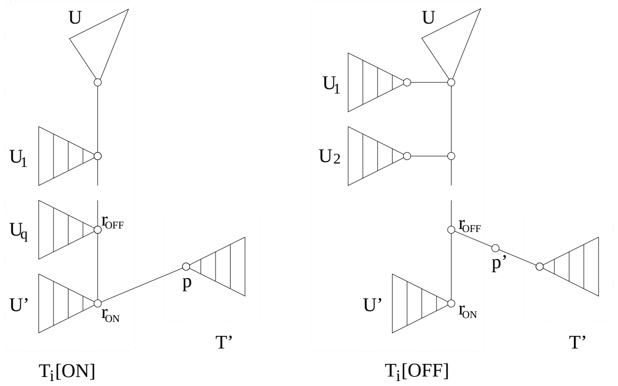

Fig. 6. T'_v and T'_v: the extension of a switch.

Suppose $\hat{T_v} \in Exp(v)$ where $\hat{T_v} = rec((T_1, \hat{r_1}), \ldots, (T_i[x], r_x), \ldots, (T_m, \hat{r_m}))$, $x \in \{ON, OFF\}$. Without loss of generality, suppose $x = ON$. Since $\hat{T_v}$ and T_v are distinct, but $(T_i[ON], r_{ON})$ is a component of T_v and $\hat{T_v}$, it must be that $r_j \neq \hat{r_j}$, for some $j \neq i$. But since $r_j = r'_j$ by the Switch Joining Lemma, it must be that $\hat{r_j} \neq r'_j$. The Switch Joining Lemma implies that $(T_i[ON], r_{ON})$ must be a component of $\hat{T_v}$ and T'_v — a contradiction. We conclude that $Exp(v)$ is the switch $\{T_v, T'_v\}$.

Case: There are at least two switches among $Exp(v_1)$, $Exp(v_2)$, …, $Exp(v_m)$.

If $|Exp(v)| = 1$ then we are done, otherwise, suppose $T_v, T'_v \in Exp(v)$ and are distinct. If T_v and T'_v differ in only one component then either the Joining Lemma or Switch Joining Lemma may be applied as in the above two cases. Otherwise, $Exp(v_i)$ is a (p, k)-switch $T_i = (U, U_1, U_2, \ldots, U_p, U', r_{ON}, r_{OFF})$ and $Exp(v_j)$ is a (q, l)-switch $T_j = (W, W_1, W_2, \ldots, W_q, W', s_{ON}, s_{OFF})$ such that $T_v = rec((T_1, r_1)$, …, $(T_i[x], r_x)$, …, $(T_j[y], s_y)$, …, $(T_m, r_m))$ and $T'_v = rec((T_1, r'_1)$, …, $(T_i[x'], r_{x'})$, …, $(T_j[y'], s_{y'})$, …, $(T_m, r'_m))$ where $x, y, x', y' \in \{ON, OFF\}$ and $x \neq x'$, $y \neq y'$. However, this contradicts the Switch Polarization Lemma. □

This article was processed using the LaTeX macro package with LLNCS style